CLAIMS

1. An apparatus comprising:

at least one processor;

a memory coupled to the at least one processor;

a database table residing in the memory; and

a cardinality estimator residing in the memory and executed by the at least one processor, the cardinality estimator estimating cardinality of an intermediate dataset that satisfies a query to the database table in a manner that accounts for data skew in the database table.

2. The apparatus of claim 1 further comprising a frequent values list residing in the memory that contains a list of values in the database table, each value having a corresponding frequency, wherein the cardinality estimator estimates the cardinality of the intermediate dataset by determining whether a frequency corresponding to a value exceeds a predetermined threshold, and if the frequency exceeds the predetermined threshold, accounting for the corresponding value, and if the frequency does not exceed the predetermined threshold, using a formula to estimate the cardinality of the intermediate dataset, the formula accounting for data skew in the database table by subtracting the frequency of all values above the predetermined threshold in the frequent value table that satisfy the query from the total number of columns in the database table.

3. The apparatus of claim 2 wherein the cardinality estimator estimates the cardinality Ca' of the intermediate dataset using the formula:

$$Ca' = P + M(1 - (1 - \frac{1}{M})^{Y})$$

where

M = Ca - (P+Q)

P = number of distinct values in the frequent values list above the predetermined threshold that satisfy the query;

Q = number of distinct values in the frequent values list above the predetermined threshold that do not satisfy the query;

Ca = cardinality of the database table;

Y = X - Fi;

X = number of rows in the intermediate dataset; and

Fi = sum of frequencies of values in the frequent values list above the predetermined threshold that satisfy the query.

4. An apparatus comprising:

at least one processor;

a memory coupled to the at least one processor;

a database table residing in the memory;

a frequent values list residing in the memory that contains a list of values in the database table, each value having a corresponding frequency; and

a cardinality estimator residing in the memory and executed by the at least one processor, the cardinality estimator estimating cardinality of the intermediate dataset using the following formula:

$$Ca' = P + M(1-(1-\frac{1}{M})^{Y})$$

where

M = Ca - (P+Q)

P = number of distinct values in the frequent values list above the predetermined threshold that satisfy the query;

Q = number of distinct values in the frequent values list above the predetermined threshold that do not satisfy the query;

Ca = cardinality of the database table;

Y = X - Fi;

X = number of rows in the intermediate dataset; and

Fi = sum of frequencies of values in the frequent values list above the predetermined threshold that satisfy the query.

5. A method for estimating cardinality of an intermediate dataset that results from processing a database query on a database table, the method comprising the steps of:

(A) evaluating the query; and

(B) estimating cardinality of the intermediate dataset using a formula that accounts for data skew in the database table.

6. The method of claim 5 wherein step (B) includes the steps of:

selecting a value in a frequent values list that contains a list of values in the database table, each value having a corresponding frequency;

if the selected value has a corresponding frequency that exceeds a predetermined threshold, incrementing the cardinality estimate by one; and

if the frequency does not exceed the predetermined threshold, using a formula to estimate the cardinality of the intermediate dataset, the formula accounting for data skew in the database table by subtracting the frequency of all values above the predetermined threshold in the frequent value table that satisfy the query from the total number of columns in the database table.

7. The method of claim 6 wherein the cardinality estimator estimates the cardinality Ca' of the intermediate dataset in step (B) using the formula:

$$Ca' = P + M(1-(1-\frac{1}{M})^{Y})$$

where

M = Ca - (P+Q)

P = number of distinct values in the frequent values list above the predetermined threshold that satisfy the query;

Q = number of distinct values in the frequent values list above the predetermined threshold that do not satisfy the query;

Ca = cardinality of the database table;

Y = X - Fi;

X = number of rows in the intermediate dataset; and

Fi = sum of frequencies of values in the frequent values list above the predetermined threshold that satisfy the query.

8. A method for estimating cardinality of an intermediate dataset that results from processing a database query on a database table, the method comprising the steps of:

(A) evaluating the query; and

(B) estimating the cardinality Ca' of the intermediate dataset using the formula:

$$Ca' = P + M(1 - (1 - \frac{1}{M})^{Y})$$

where

M = Ca - (P+Q)

P = number of distinct values in the frequent values list above the predetermined threshold that satisfy the query;

Q = number of distinct values in the frequent values list above the predetermined threshold that do not satisfy the query;

Ca = cardinality of the database table;

Y = X - Fi;

X = number of rows in the intermediate dataset; and

Fi = sum of frequencies of values in the frequent values list above the predetermined threshold that satisfy the query.

9. A program product comprising:

(A) cardinality estimator estimating cardinality of an intermediate dataset that satisfies a query to a database table in a manner that accounts for data skew in the database table; and

(B) computer-readable signal bearing media bearing the cardinality estimator.

10. The program product of claim 9 wherein the computer-readable signal bearing media comprises recordable media.

11. The program product of claim 9 wherein the computer-readable signal bearing media comprises transmission media.

12. The program product of claim 9 wherein the cardinality estimator evaluates a frequent values list that contains a list of values in the database table, each value having a corresponding frequency, wherein the cardinality estimator estimates the cardinality of the intermediate dataset by determining whether a frequency corresponding to a value exceeds a predetermined threshold, and if the frequency exceeds the predetermined threshold, accounting for the corresponding value, and if the frequency does not exceed the predetermined threshold, using a formula to estimate the cardinality of the intermediate dataset, the formula accounting for data skew in the database table by subtracting the frequency of all values above the predetermined threshold in the frequent value table that satisfy the query from the total number of columns in the database table.

13. The program product of claim 12 wherein the cardinality estimator estimates the cardinality Ca' of the intermediate dataset using the formula:

$$Ca' = P + M(1-(1-\frac{1}{M})^{Y})$$

where

M = Ca - (P+Q)

P = number of distinct values in the frequent values list above the predetermined threshold that satisfy the query;

Q = number of distinct values in the frequent values list above the predetermined threshold that do not satisfy the query;

Ca = cardinality of the database table;

Y = X - Fi;

X = number of rows in the intermediate dataset; and

Fi = sum of frequencies of values in the frequent values list above the predetermined threshold that satisfy the query.

14. A program product comprising:

(A) a cardinality estimator that estimates cardinality of the intermediate using the following formula:

$$Ca' = P + M(1 - (1 - \frac{1}{M})^{Y})$$

where

M = Ca - (P+Q)

P = number of distinct values in the frequent values list above the predetermined threshold that satisfy the query;

Q = number of distinct values in the frequent values list above the predetermined threshold that do not satisfy the query;

Ca = cardinality of the database table;

Y = X - Fi;

X = number of rows in the intermediate dataset; and

Fi = sum of frequencies of values in the frequent values list above the predetermined threshold that satisfy the query; and

(B) computer-readable signal bearing media bearing the cardinality estimator.

15. The program product of claim 14 wherein the computer-readable signal bearing media comprises recordable media.

16. The program product of claim 14 wherein the computer-readable signal bearing media comprises transmission media.

* * * * *